The meaning of morphomes: distributional semantics of Spanish stem alternations

Borja Herce, Marc Allassonnière-Tang

Borja Herce* and Marc Allassonnière-Tang

The meaning of morphomes: distributional semantics of Spanish stem alternations

https://doi.org/10.1515/lingvan-2023-0010
Received January 11, 2023; accepted September 11, 2023; published online March 20, 2024

Abstract: Romance stem alternations have been argued to represent exclusively morphological objects (or "morphomes") independent from semantic and syntactic categories. This conclusion has been based on feature-value analyses of the inflected forms, and definitions of natural classes that are theoretically driven and about which no consensus exists. Individual examples of morphomes are thus frequently challenged, while their autonomously morphological nature has never been tested quantitatively or experimentally. This is the purpose of the present study. We use context-based embeddings to explore the semantic profile of Spanish verb stem alternations. At the paradigmatic level, our findings suggest that Spanish morphomes' cells are characterized by significantly above-chance distributional-semantic similarity. At the lexical level, similarly, verbs that show more similar patterns of alternation have also been found to be closer in meaning. Both of these findings suggest that these structures may have an extramorphological function. Using gradient distributional-semantic similarity offers a way to objectively assess the degree of (un)naturalness of a set of forms and meanings, something which has been lacking from most discussions on the structure of features and the architecture of paradigms.

Keywords: morphome; Spanish; morphology; semantics; stem alternations; word embeddings

1 Introduction

Stem alternations in Romance have been investigated extensively in the morphological literature over the last decades (e.g. Esher 2017; Herce 2022b; Maiden 2005, 2018; Malkiel 1966). They have been the central object of analysis in discussions concerning the autonomous morphology hypothesis (Aronoff 1994; Esher 2012; Herce 2023; Luís and Bermúdez-Otero 2016; O'Neill 2014), according to which some grammatical structures are exclusively morphological and do not match any syntactic or semantic domain.[1] Romance stem alternations have been argued to be "morphomes" (i.e. autonomously morphological structures), in opposition to "morphemes" which do match extramorphological values or categories. A commonly used succinct definition of the phenomenon is that the morphome is "a systematic morphological syncretism which does not define a (syntactically or semantically) natural class." (Trommer 2016: 60).

The problem with a definition like this, and by extension with the way morphomes have been usually identified and discussed in the literature, is that we are nowhere close in the field to a consensus on what exactly should count as a "natural" class. On the basis of one and the same paradigm, morphologists with different (theoretical) inclinations may posit different features with different values and architectures. Hence, they will often derive a different number of natural syntactic/semantic classes extending over different sets of word or paradigm cells (e.g. Aalberse 2007; Harbour 2016; Wyngaerd 2018). What is missing, in order to objectify the identification of these domains, is a purely empirical approach to the exploration of syntactic/semantic similarities and differences.

1 Independence from phonological domains has also been often discussed (e.g. Anderson 2013; Herce 2020b; Maiden 2017) but will not be addressed in this paper.

***Corresponding author: Borja Herce**, Department of Comparative Language Science, University of Zurich, Zurich, Switzerland, E-mail: borja.hercecalleja@uzh.ch. https://orcid.org/0000-0002-9493-6656
Marc Allassonnière-Tang, Lab Eco-Anthropology UMR 7206, MNHN, Paris, France, E-mail: marc.allassonniere-tang@mnhn.fr. https://orcid.org/0000-0002-9057-642X

The distributional-semantic tenet that "you shall know a word by the company it keeps" (Firth 1957: 11) offers the way to do precisely this. Looking at the relative similarity of the context of occurrence of different words (e.g. different lemmas like DRIVE, LEAD, or ASK or different word forms of the same lemma like *drives*, *drove*, or *driven*) offers a way around these categorical and theory-dependent natural classes and thus has the potential to overcome the single greatest challenge in diagnosing morphomicity: the lack of positive diagnostics, as Koontz-Garboden (2016) puts it. In addition, because the natural versus unnatural axis is most likely not a simple dichotomy (Andersen 2008; Herce 2020a; Saldana et al. 2022; Smith 2013), this approach also offers the opportunity to modulate this notion, and approach and measure it in a finer-grained way. This is the goal of the present paper methodologically.

The challenge is considerable, however, and cannot be approached holistically. Because, as mentioned earlier, Romance stem alternations constitute the most frequently discussed morphomic object in the literature, they constitute an ideal litmus test for the autonomous morphology hypothesis. Different Romance varieties can differ quite substantially in their stem alternation patterns, and the present paper cannot possibly explore all of them. For reasons of data availability and because of its conservative nature when it comes to stem alternations, we focus on Spanish. The distributional-semantic profile of verbal stem alternation patterns and conjugations in the language will be explored at the paradigmatic and lexical levels to assess just how autonomous these are from syntax and semantics.

The structure of the paper is the following. Section 2 Introduces the morphological objects (i.e. mostly stem alternation patterns, but also conjugations) that this paper explores. Section 3 describes the dataset and methodology employed to measure their distributional-semantic profile. Section 4 presents the results and Section 5 discusses their interpretation. Section 6 contains conclusions and avenues for future research.

2 Spanish morphomes and conjugations

Spanish (and Romance) morphomes constitute abstract verbal paradigmatic domains over which a stem alternant frequently differs from the one found elsewhere in the paradigm. Four such domains have been identified in the literature. These have received the labels N(-pattern), L(-pattern), F(UÈC) and P(YTA) (see Maiden (2018) for the story behind these labels).

The paradigms of some "irregular" Spanish verbs can be used to illustrate the paradigmatic domains of these morphomes. The paradigm of *salir* 'come' in Table 1 shows the extension of L (stem alternant *salg-* in light gray) and F (stem alternant *saldr-* in dark gray). The former spans over the first person singular present indicative and all cells of the present subjunctive. The latter expands over all the person/number cells of the future and the conditional. In the paradigm of *querer* 'want' in Table 2, we show the domains of the N and P morphomes. The former (stem alternant *quier-* in light gray) spreads over the singular and third person plural of the present indicative and present subjunctive, and to the second person singular imperative (*quiere*, not shown in the table). The latter (stem alternant *quis-* in dark gray) spreads over the preterite and imperfect subjunctive tenses shown in Table 2, as well as to two other tenses not shown in the table.[2]

The historical origin of these paradigmatic structures is heterogeneous. P continues the distinct perfectum stem that many Latin second and third conjugation verbs had to signal aspect. N and L emerged from sound changes (unstressed vowel mergers and palatalizations respectively) that occurred after Classical Latin but before the breakup of (Continental) Romance. F, in turn, derived from the emergence of new synthetic tenses in (Western) Romance from former periphrases involving the infinitive and forms of the verb 'have'. Regardless of their origin, the claim common to all of these structures (particularly with respect to N, L, and P) is that they synchronically involve a heterogeneous set of cells or tenses that have no syntactic/semantic property that

2 One (with forms *quisiese*, *quisieses*, *quisiese*, *quisiésemos*, *quisieseis*, *quisiesen*) is generally considered to be synonymous with the imperfect subjunctive (i.e. with *quisiera*, *quisieras*, etc.). The other is the future subjunctive (with forms *quisiere*, *quisieres*, *quisiere*, *quisiéremos*, *quisiereis*, *quisieren*), which is hardly ever used in the modern language.

Table 1: Paradigm of *salir* 'come', illustrating L (light gray) and F (dark gray) morphomes.

	Present indicative	Present subjunctive	Imperfect indicative	Preterite	Imperfect subjunctive	Future	Conditional
1SG	*salgo*	*salga*	*salía*	*salí*	*saliera*	*saldré*	*saldría*
2SG	*sales*	*salgas*	*salías*	*saliste*	*salieras*	*saldrás*	*saldrías*
3SG	*sale*	*salga*	*salía*	*salió*	*saliera*	*saldrá*	*saldría*
1PL	*salimos*	*salgamos*	*salíamos*	*salimos*	*saliéramos*	*saldremos*	*saldríamos*
2PL	*salís*	*salgáis*	*salíais*	*salisteis*	*salierais*	*saldréis*	*saldríais*
3PL	*salen*	*salgan*	*salían*	*salieron*	*salieran*	*saldrán*	*saldrían*

Table 2: Paradigm of *querer*, 'want' showing N (light gray) and P (dark gray) morphomes.

	Present indicative	Present subjunctive	Imperfect indicative	Preterite	Imperfect subjunctive	Future	Conditional
1SG	*quiero*	*quiera*	*quería*	*quise*	*quisiera*	*querré*	*querría*
2SG	*quieres*	*quieras*	*querías*	*quisiste*	*quisieras*	*querrás*	*querrías*
3SG	*quiere*	*quiera*	*quería*	*quiso*	*quisiera*	*querrá*	*querría*
1PL	*queremos*	*queramos*	*queríamos*	*quisimos*	*quisiéramos*	*querremos*	*querríamos*
2PL	*queréis*	*queráis*	*queríais*	*quisisteis*	*quisierais*	*querréis*	*querríais*
3PL	*quieren*	*quieran*	*querían*	*quisieron*	*quisieran*	*querrán*	*querrían*

distinguishes them from others, and hence that they constitute unnatural classes. These have proven to be, however, quite resilient and semi-productive categories in Romance (see Maiden (2018) for a more detailed exposition of their diachrony).

In Spanish, and Romance more generally, the problem of diagnosing (un)naturalness is most acute with respect to different tenses (i.e. present indicative, present subjunctive, imperfect indicative, etc.), as these cannot be easily arranged into a set of features with mutually exclusive values. As with grammatical cases (like "nominative" or "ablative"), these labels correspond to language-specific morphological entities (Haspelmath 2010); a set of person/number endings (e.g. *-í*, *-iste*, *-ió*, *-imos*, *-isteis*, *-ieron*) or a morpheme (e.g. *-ía-*, *-ra-*, etc.) with a complex range of (sometimes only loosely connected) syntactic and semantic uses which are unlikely to be successfully captured by a low number of abstract features and values.

Even with regard to seemingly easier features, like person and number, uncertainties persist. It is unclear, for example, if some values of person should be deemed more similar to each other. If we believe that first and second person have more in common, we could set a feature ±speech-act-participant. If we believe second and third person have more in common we can speak of ±speaker. Even after we make a decision on this, it still needs to be decided whether the same hierarchy should apply across number values. In a language without clusivity like Spanish, first person plural very often includes the addressee, so it could plausibly be judged more similar to second person plural than first person singular is to second person singular. Some or perhaps most of these uncertainties and disagreements do not really derive from a lack of data, or from untested hypotheses, but reflect a more fundamental problem with the way syntactic/semantic structure has been traditionally analyzed and formalized, as a very low-dimensional vector of "same" versus "different" values. A large-dimensional replicable empirical approach could be the way out of this quagmire, allowing us to assess the degree of (un)naturalness of different sets of cells or forms (more) objectively.

The same issues that we have just identified in the analysis of the syntactic and semantic structure of paradigmatically related forms arise even more prominently with respect to the structure of the lexicon. Different verbs (e.g. *salir* vs. *tener* in Tables 1 and 2) belong in Spanish to different conjugations, that is, they take a different set of endings (compare infinitive *sal-ir* of the third conjugation with *ten-er* of the second). These conjugations have also been argued to constitute purely morphological entities (they have been recently labeled "rhizomorphomes" by Round (2015)) because verbs from the same conjugation (e.g. *venir* 'come', *herir* 'hurt', *vivir* 'live', *salir* 'exit', *abrir*

'open', etc.) are said not to be characterized by any common semantic/syntactic property, but merely by their shared inflections. The semantic and syntactic properties of lexemes, however, differ along so many different aspects that we are most unlikely to arrive at a small set of features (e.g. ±volitional, ±transitive, ±beneficial, ±movement) that successfully characterize all lexemes in a language and can be used to assess their relative semantic/syntactic similarity, and hence whether a subset of them constitutes a natural class. Because of the greater size and relative disorderliness of the lexicon compared to the average inflectional paradigm, few attempts have been made to pursue this. Although attempts have been made at small and comparatively orderly semantic fields such as kinship terms (Allen 2008; Pericliev and Valdés-Pérez 1998; Radcliffe-Brown 1941), lexical similarities have been more often investigated holistically, via speaker judgments, or via the similarity of the contexts where different words occur (Miller and Charles 1991). The latter approach has become much faster and robust with the increase of computing power and the advent of methods like word2vec, to assess the contextual (i.e. embedding) similarities of different words. Adopting this approach, the main research questions of the present paper will be:

(1) How (un)natural are the sets of cells over which the L, N, P, and F stem alternants span in the verbal paradigm of Spanish?
(2) How semantically and syntactically homogeneous are verbs with the same or similar patterns of stem alternation?
(3) How similar are verbs from the same conjugation?

Overall, then, the goal is to find out whether there are distributional-semantic correlates of structures that have been characterized as "morphomic".

3 Data and methods

The working hypothesis of distributional semantics is that the context of use of a word determines or captures its meaning. Words with similar meanings (e.g. *murder*, *kill*, *assassinate*) will tend to occur in similar syntactic and semantic environments. In terms of syntax, in the specific example given, these words (verbs) will tend to occur after nouns or noun phrases (denoting animate agents), and also before nouns or noun phrases (denoting animate patients). With respect to semantics, these words will tend to be found in sentences dealing with violence and crime and its concomitants: firearms, knives, blood, police, judges, trials, and so on. On the other hand, words with very different meanings (e.g. *murder*, *smart*, *ago*) will tend to occur in very different syntactic and semantic environments. Thus, the similarity of the contexts where two words occur can provide a measure of the overall syntactic and semantic similarity of the words.

There is an abundant literature that has successfully followed this approach. Because, as mentioned in Section 1, the structure of the lexicon is regarded as much more complex and unpredictable than that of the average inflectional paradigm (Bonami and Paperno 2018), this approach has been used more often to explore the former, and derivational processes and families (Huyghe and Wauquier 2020), rather than the latter. However, there is nothing in the method that makes it unsuitable to investigate the structure of inflectional paradigms (Kirschenbaum 2021). On the contrary, due to the feature and value structure-related analytical uncertainties we described in Section 2, approaching inflectional paradigmatic structure from a distributional-semantic perspective may constitute a more empirical and replicable route for finding abstract structural categories and principles (Chuang et al. 2022; Nikolaev et al. 2022).

We use the Spanish Billion Words Corpus and Embeddings (SBWCE) linguistic resource from Cardellino (2019), which is the largest available Spanish corpus. It is freely usable and downloadable from https://crscardellino.ar/SBWCE/.[3] This is a corpus with 1,420,665,810 words, in which 1,000,653 different words occur five times or more. For

3 We performed the same analyses we describe here on the Spanish language subcorpus of Universal Dependencies De Marneffe et al. (2021), but found SBWCE preferable. Although this corpus is annotated, the annotation is not always reliable and did not compensate for the shortcomings derived from its much smaller size (over 1,000 times smaller), which left many word forms of even very high frequency lemmas unattested or insufficiently attested.

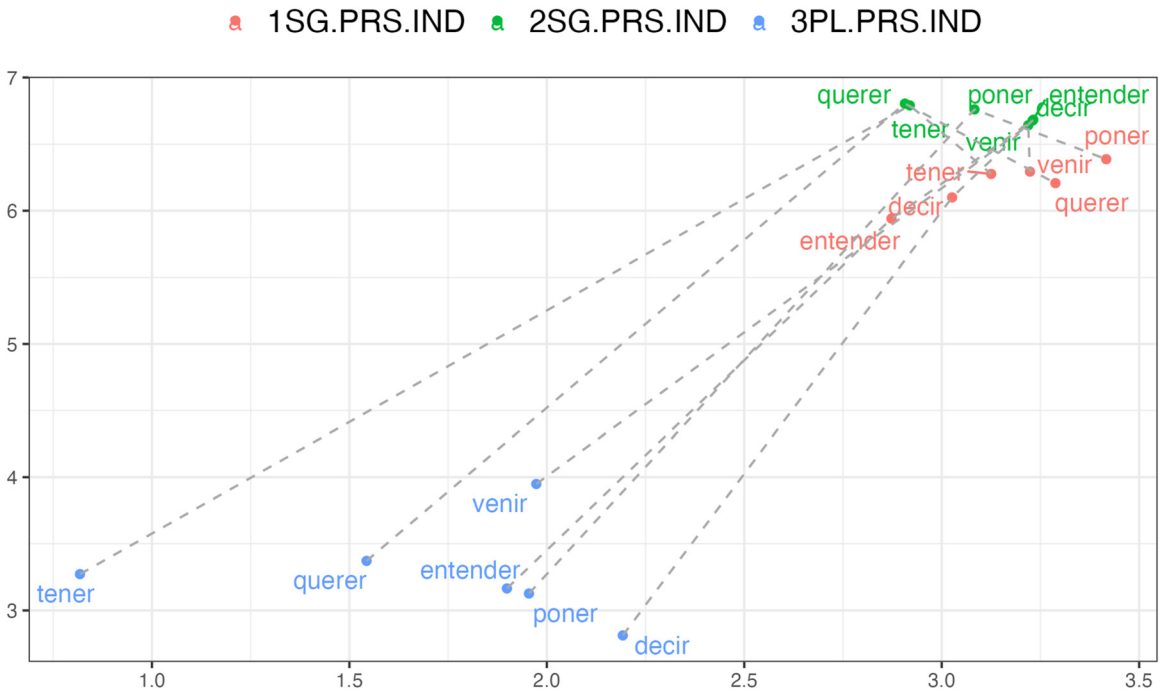

Figure 1: An example of the embedding distance between three forms from six verbs: *tener* 'have', *querer* 'want', *venir* 'come', *entender* 'understand', *poner* 'put', and *decir* 'say'. All forms represent present indicative cases; those in red represent first person singular, those in green represent second person singular, and those in blue represent third person plural.

these, pre-trained embeddings consist of 300 dimensions, which we could use for our analyses. To avoid data sparsity and to keep computational processing times in check, we focused on the 200 most frequent verbal lemmas, whose inflected word forms (identified from Unimorph; McCarthy et al. (2020)) total 12,054. After discarding homophones and homographs (e.g. *sal* 'salt/exit.2SG.IMP') and identifying syncretic forms (e.g. *corre* 'run.2SG.IMP/run.3SG.PRS.IND'), we explored the similarities between the word embeddings associated to all remaining forms.[4]

As an example, Figure 1 shows the relative similarity of *tengo*, *tienes*, *tienen* (first person singular, second person singular, and third person plural present indicative respectively of the verb *tener* 'have'), and the equivalent forms of other high-frequency verbs. The embedding distances between the different forms are roughly parallel to the similarity between the forms' values as usually described by morphologists. First person singular present indicative (in red) and second person singular present indicative (in green) have more in common with each other (a value "singular") than either of them have with the third person plural present indicative (in blue). The former pair of cells, thus, constitute a more natural class than the latter pair. In this way, by measuring the relative embedding distances between different word forms and between different lemmas, Section 4 will present an answer to the research questions outlined in Section 2.

4 Results

4.1 Inflectional-semantic naturalness of morphomes

The sets of paradigm cells that may share a special stem in Spanish (see Tables 1 and 2) have been argued to be morphomic (i.e. unnatural classes of cells with respect to their syntactic and semantic profile). Unlike most

4 The following abbreviations are used in this paper: 1/2/3 first/second/third person; IMP imperative; IND indicative; PRS present; SG singular; PL plural.

current formalizations suggest, however, naturalness and syntactic/semantic affinity should probably not be regarded as dichotomous properties but rather as gradient dimensions (cf. Baayen et al. (2019)). This means that asking whether any two cells constitute a natural class or not might be a nonsensical question, akin to asking whether Berlin and Paris are "close". Only in comparison to other pairs of cells (or cities) does it make sense to say whether they are "closer to" or "farther away" from each other. The question we will be asking (Research Question 1 in Section 2), is thus how (un)natural the sets of paradigm cells are that partake in the same stem alternation patterns in Spanish verbs. Figure 2 shows the embedding distance (average cosine similarity) of pairs of cells from the same morphomic domain (in blue) compared with the averages of 1,000 random samples of the same size (the gray histogram, with 90 % confidence intervals in red). We pursue a bootstrapping approach to significance because this sidesteps the lack of independence of distance measures between repeated forms, and because it was found to be the most conservative among the options we explored (Wilcoxon signed rank test with Bonferroni correction, and t test).

The blue line in L shows the average similarity between L cells (within and across lemmas, i.e. *salgo*, *salgas* as well as *salgo*, *quieras*), while the histogram represents the average similarities of 1,000 random samples of sets of cells of the same size (in this case, sets of six).[5] The same applies to the other domains for stem allomorphy, that is, N, P, and F.

The results reveal that paradigm cells from the morphological domains L and P, tend to be characterized by significantly greater syntactic/semantic similarity than a randomly selected set of cells (i.e. these forms are more similar than in 95 % of samples of the same size). F, in turn, borders significance but falls slightly short of it, while N cells appear not to be significantly more coherent than if they had been selected at random. If our notion of naturalness is gradient, then we can say that most Spanish morphomes lean more towards extreme naturalness than towards extreme unnaturalness. Most natural would be P, a "TAM morphome" (see Smith (2013)) which includes all person/number forms of those tenses that used to be perfective in Classical Latin. Next comes L, where all cells except one share identical TAM values. Least natural (and the only one fairly describable as morphomic under this approach) is N. This unnaturalness might derive from the fact that its cells span all moods in the language: indicative, subjunctive, and imperative.

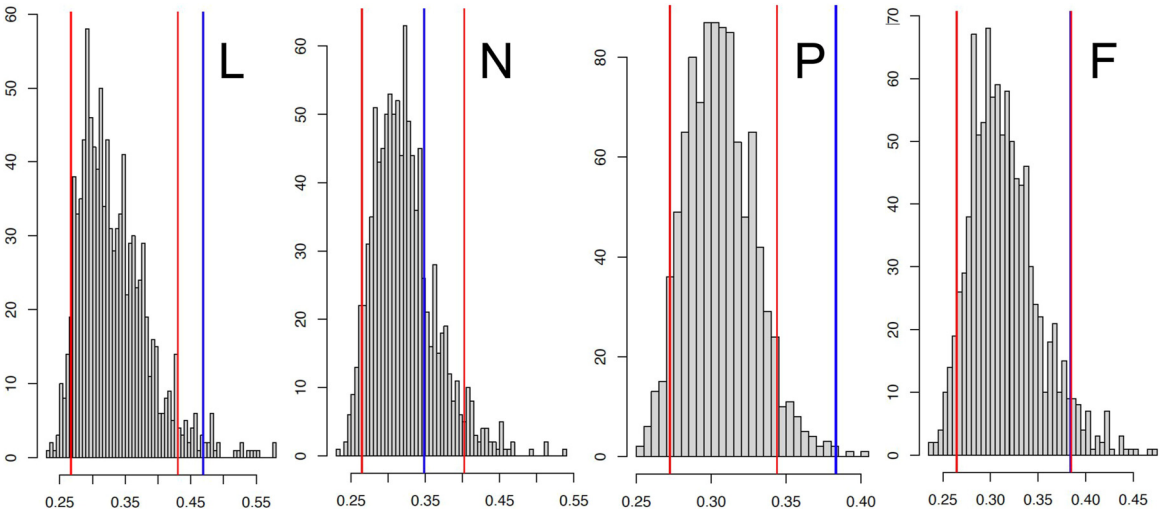

Figure 2: Embedding distance of L, N, P, and F cells compared to other cells. The blue line shows the average embedding distance (cosine similarities) of pairs of cells from the same morphomic domain. The red lines indicate the 90 % confidence interval of the averages of 1,000 random samples of the same size.

5 We define cells here, like Boyé and Schalchli (2019), with reference to distinct word forms. Thus, L expands over six word forms (e.g. *salgo*, *salga*, *salgas*, *salgamos*, *salgáis*, and *salgan* in Table 1).

4.2 Lexical-semantic naturalness of morphomes

Beyond the distributional-semantic similarity of their different inflectional values, an extramorphological role for Romance morphomes could be plausibly sought at the lexical-semantic level. That is, a morphological entity could receive functional justification either via the grammatical values it appears in, or via the lexical meanings with which it appears. We would like to know, therefore, if verbs with identical or more similar patterns of stem alternation are more similar in their lexical semantics. As Figure 3 shows, we found that verbs which are morphomically identical and share multiple patterns of alternation (e.g. *decir* 'say' and *hacer* 'do' are identical in that they display L, P, and F alternation patterns, but no N) are significantly more similar semantically.

Verbs which have only one shared pattern of stem alternation (e.g. *decir* 'say' and *parecer* 'seem' share L) are less similar semantically (slightly below statistical significance), but still more so than those than verbs that do not share any pattern of alternation (e.g. *decir* 'say' and *perder* 'lose', where the former has L, P, and F, and the latter N). Verbs which have no stem alternation whatsoever (e.g. *vivir* 'live' and *amar* 'love', which are both "regular" verbs with an unchanging stem) appear to be, surprisingly, significantly more dissimilar than randomly selected verbs.

4.3 Lexical-semantic naturalness of conjugations

Inflection classes (i.e. declensions and conjugations) are another type of morphological entity that has been claimed to be autonomously morphological. In Spanish, verbs usually belong to one of three different conjugations.[6] Verbs from the same conjugation share their inflectional endings but are supposed to share no meaning. Our Research Question 3 was aimed at assessing whether or not this is the case. According to our results (displayed in Figure 4), this seems to be largely true.

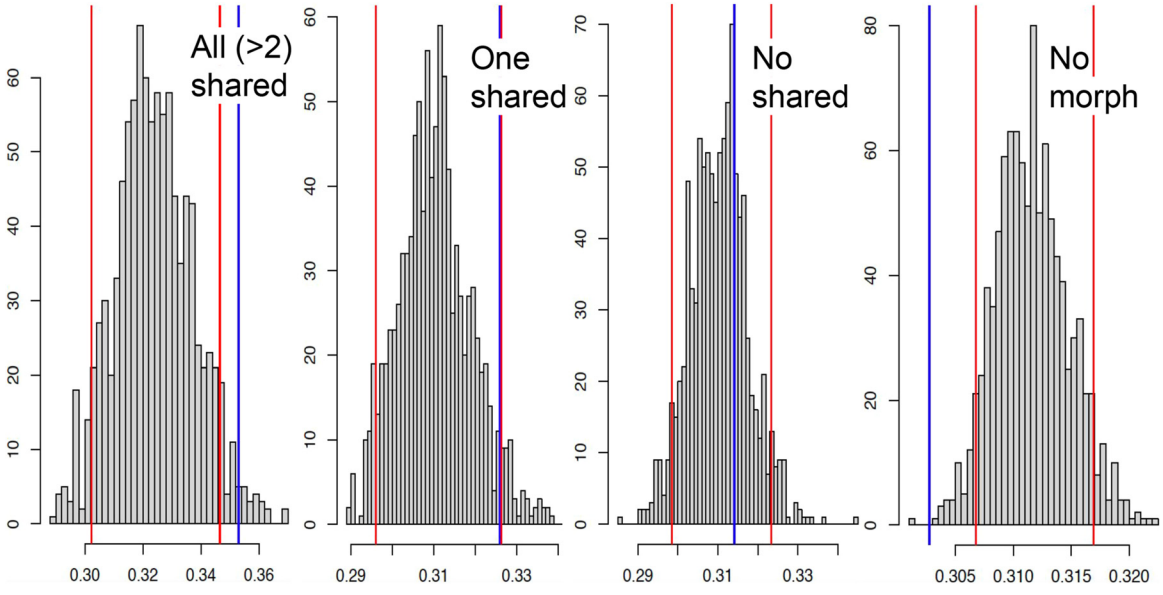

Figure 3: Embedding distance between verbs sharing multiple (all) stem alternation patterns (morphomes), sharing one alternation pattern, and sharing no patterns, and between verbs having no stem alternations. The blue line shows the average embedding distance (cosine similarities) of pairs of cells from each category. The red lines indicate the 90 % confidence interval of the averages of 1,000 random samples of the same size.

6 Some verbs, so-called heteroclites, may belong to different conjugations in different parts of their paradigm. For example, *dar* behaves as a first conjugation verb in most of the paradigm (e.g. *d-ar*, *d-amos*, *d-es*, *d-aba* parallel to *am-ar*, *am-amos*, *am-es*, *am-aba*) but as a non-first conjugation verb in P forms (e.g. *d-iste*, *d-ieron* parallel to *viv-iste*, *viv-ieron*). These (few) verbs have been excluded due to their mixed conjugational affiliation.

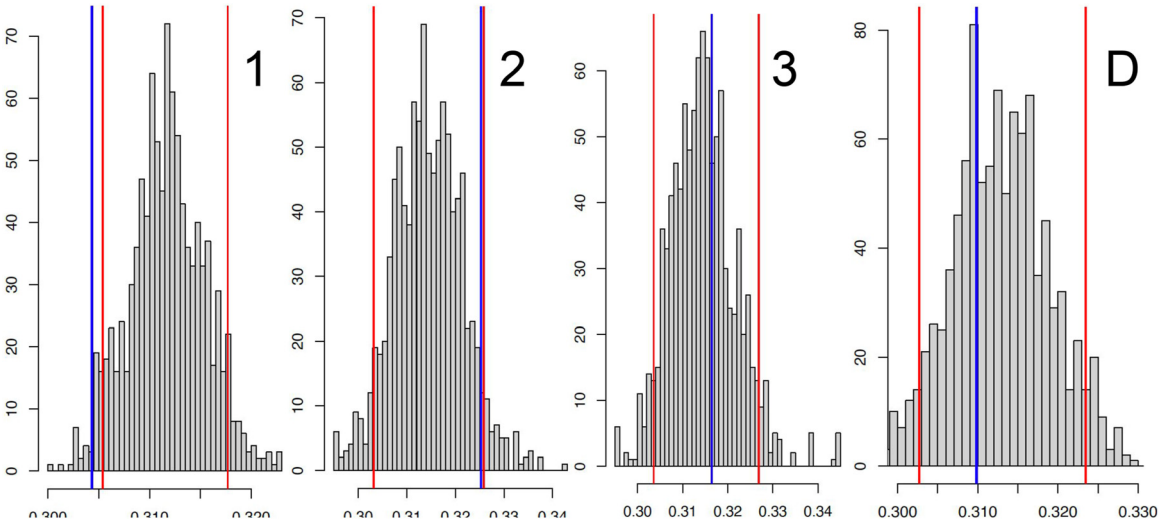

Figure 4: Embedding distance between verbs of different Spanish conjugations. The blue line shows the average embedding distance (cosine similarities) of pairs of cells from verbs from each conjugation type (first, second, third, and different conjugations). The red lines indicates the 90 % confidence interval of the averages of 1,000 random samples of the same size.

Although second conjugation verbs (labeled 2) do come extremely close to being significantly more similar on average than a same-sized random sample of verbs, third conjugation verbs (labeled 3) and verbs from different conjugations (labeled D) are not significantly different from randomly selected verbs in their distributional-semantic similarity. First conjugation verbs (labeled 1) in turn, appear to be significantly less similar than randomly selected samples of verbs. This finding might be related to the greater dissimilarity of verbs without stem alternations. Both first conjugation and absence of alternation constitute the most token-frequent and productive morphological configurations in Spanish and could thus be regarded as defaults.

4.4 A confound? Shared roots

Beyond morphomes and inflection classes, another aspect that has sometimes been argued to be morphomic is root or morph sharing (see Aronoff's [1994: 28] and Round's [2015] notion of the meromorphome). That is, verbs across Romance may be based upon the same root but not share any immediately apparent semantic trait or affinity. This is the case, for example, with *explicar* 'explain', *replicar* 'answer', *complicar* 'make difficult', *aplicar* 'apply', *suplicar* 'beg', all of which share a (meaningless?) root *plicar*. Note that this morphological property interacts very significantly with the question we asked in considering Figure 3. Verbs which are based upon the same root usually share all their stem alternation quirks. Thus, verbs like *tener* 'have', *contener* 'contain', *sostener* 'hold', and *detener* 'stop/arrest', or *poner* 'put', *componer*, 'compose' and *imponer* 'impose' may share their stem alternation patterns (L, P, and F in these cases) at least in part due to their shared root, which might plausibly be associated to shared meaning as well. Hence we need to compare whether, or to what extent, root identity is associated with meaning similarity (Figure 5).

The results show that root sharing is not associated with distributional-semantic similarity. Root sharing in Spanish verbs is thus ratified as a purely morphological property and is hence unlikely to be the factor driving the greater similarity of verbs with identical or similar patterns of stem alternation that we reported in Figure 3.

5 Discussion

The results described in Section 4 can be interpreted in different ways. They can be understood to cast doubt, or at least demand a more moderate and finer-grained stance, on the autonomously morphological nature of Spanish

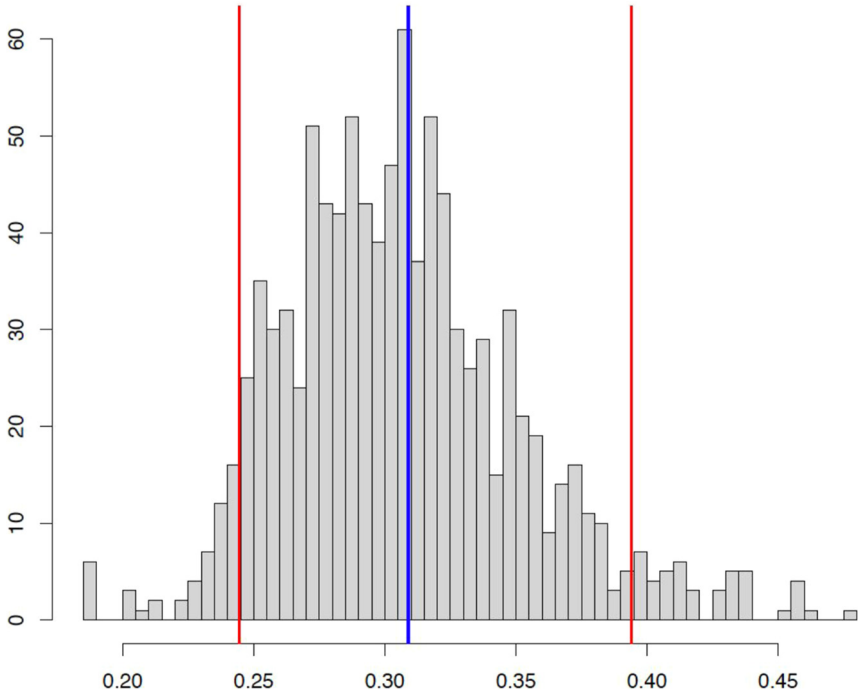

Figure 5: Embedding distance between verbs containing the same root. The blue line shows the average embedding distance (cosine similarities) of pairs of cells sharing the same root. The red lines indicates the 90 % confidence interval of the averages of 1,000 random samples of the same size.

(and by extension Romance) verb stem alternation patterns. At the paradigmatic level, the results in Figure 2 show that the word forms that belong to these morphological domains tend to be characterized by a higher degree of embedding similarity than comparable (i.e. same-sized) sets of cells. This is an observation which might have been made impressionistically. Within the traditional tabular representations of paradigms, Romance morphomic domains do not seem to involve semantically and syntactically random sets of cells (see Figures 1 and 2), but more orderly and "contiguous" swaths of the paradigm, such as whole tenses. We understand this behavior as part and parcel of the paradigmatic distribution of Spanish morphomes and hence of their degree of (un) naturalness. To anyone interested in different questions, however, our data will also allow the exploration of the embedding similarity of different tenses (see Figure 6, where we compare third person plural cells).

Transcending the dichotomy between morphemic and morphomic, the approach we employ here makes it possible to quantify the degree of syntactico-semantic homogeneity of the respective domains. This is what has allowed us to identify the P morphome as the most natural and the N morphome as the least natural one in Spanish. This scale of naturalness (P > L > F > N) differs from the one arrived at by Smith (2013), according to which P and F, because they only involve inherent inflectional categories (Booij 1996) (i.e. TAM values), should count as more functionally coherent than those like L and N which involve also contextual inflectional categories (i.e. person and number).

At the lexical level, the results in Figure 3 suggest that verbs sharing morphomes also tend to be closer semantically than verbs with different patterns of alternation or with no alternations. This is understandable if we consider that the semantic similarity of two verbs might plausibly promote morphological similarity; for example, by making analogical changes more probable when the target and the model of an analogical change are close semantically and syntactically. The L stem alternation in Spanish *hacer*, for example, must have come about analogically, with the (etymological) pattern of alternation in *decir* (one of) the main suspect(s) for having acted as the model in a four-part analogical change (i.e. /diθe/:/digo/, /aθe/:/?/).[7]

Reasons for change can be many, and pressures nondeterministic, but it is reasonable to assume that syntactically or semantically similar verbs (also phonologically, and morphologically similar ones) should

7 Note that although modern Spanish shows third person singular present indicative /aθe/ and first person singular present indicative /ago/, the etymologically expected forms would be /aθe/ and /aθo/ respectively.

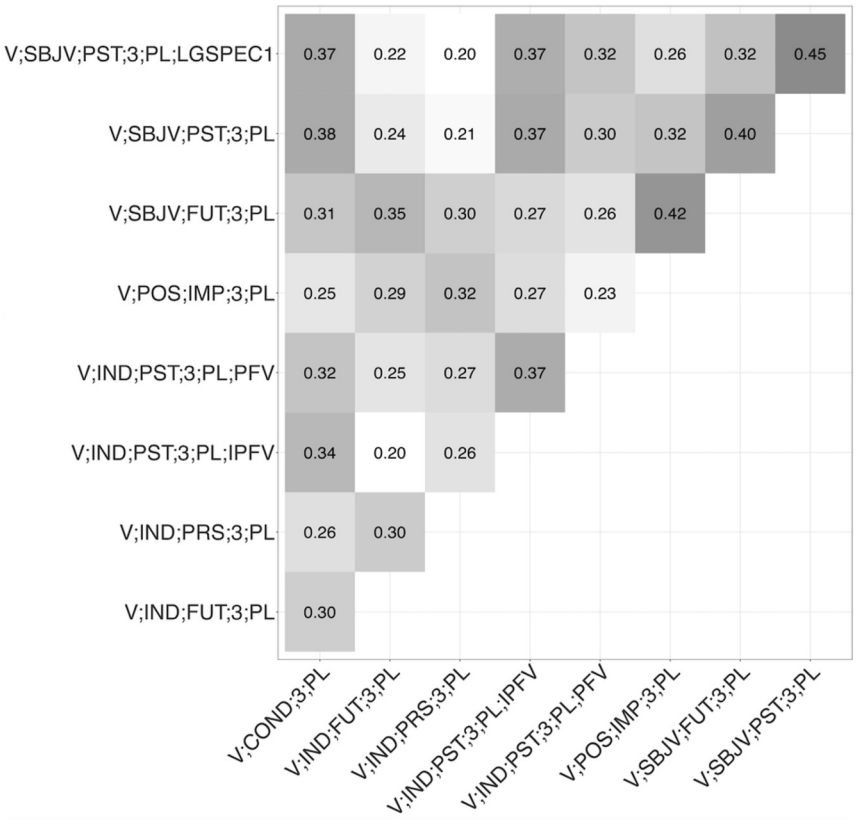

Figure 6: Embedding similarity of third person plural cells from different tenses.

influence each other to a greater degree than syntactically or semantically more distant verbs, so that over time they will tend to gravitate towards the same phonological and morphological properties. The ability of semantic similarity to generate morphological similarity is widely acknowledged, being witnessed most clearly in cases of so-called lexical contamination. Thus, closeness of meaning is the main explanation for countless one-off changes like Russian *devyat* < *nevyat 'nine' (under the influence of *desyat* 'ten'), Late Latin *sinɛstru < *sinestru 'left' (under the influence of *dɛstru 'right'), and for the intrusion of an/l/segment into many Romance reflexes of Latin *possum* 'can', under the influence of *vōlo* 'want' (Maiden 2004: 236).

Of course, and although this would go against the principle of morphology-free syntax (Zwicky 1996: 301), we cannot exclude the idea that influence might be bidirectional. Shared form and similar or identical patterns of stem alternation might promote distributional-semantic similarity as well. Research on (morphological) priming (Bentin and Feldman 1990; Veríssimo and Clahsen 2009) has established that morphological similarity can lead to co-activation, and hence to an increased probability of morphologically similar words being repeated very close to each other and in similar frames. Thus, the fact that, for example, *poder* 'can' and *querer* 'want' share all their patterns of alternation in Spanish (N, P, and F) might make it easier for them to prime each other to some extent, thus increasing the likelihood that they will be used in the vicinity of each other (e.g. in set phrases like *querer es poder* 'where there's a will there's a way') and in similar contexts (see also the literature on categorization and language acquisition based on "frequent frames"; Mintz 2003; Chemla et al. 2009).

Although Spanish stem alternations have been found here to be significantly associated with above-chance lexical-semantic and inflectional-semantic coherence, these findings need to be put into perspective. First, the distributional-semantic affinities we detect here are subtle, generally much more so than the ones associated with the classes, categories, or values that linguists generally discuss and label (see e.g. Figure 7), which are often more similar than all random samples, rather than merely 95 % of them.

At the same time, semantic/syntactic associations have not been found for other allegedly morphomic phenomena in Spanish. Thus, we found that, at least among the very-high-frequency verbs we analyze, sharing a

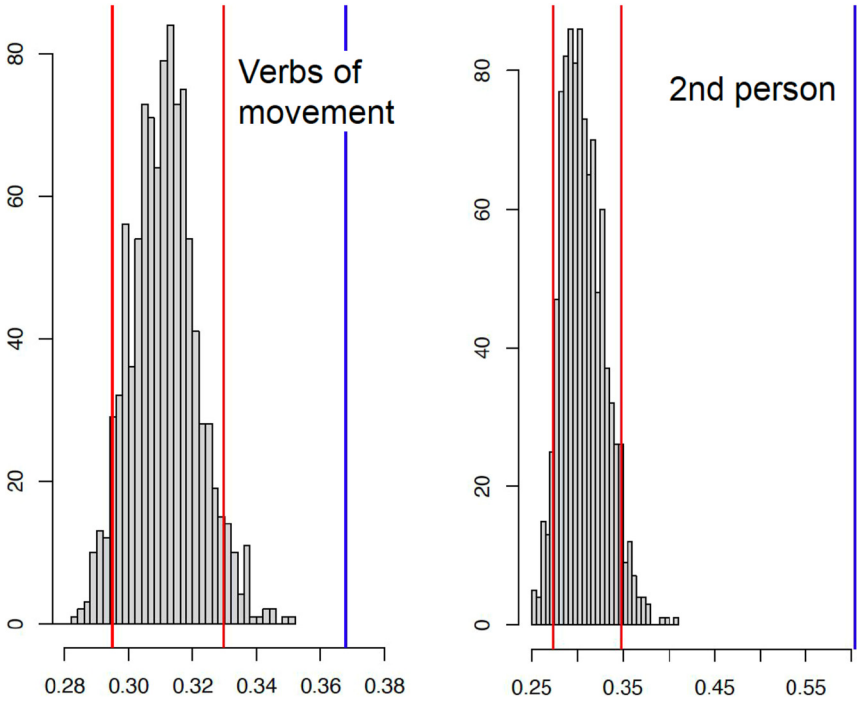

Figure 7: Two extremely natural classes in lexicon (*left*) and paradigm (*right*).

root is (perhaps surprisingly) not associated with greater distributional-semantic similarity (see Figure 5).[8] Similarly, little to no semantic correlate was found for Spanish conjugations (see Figure 4). Thus, with the possible exception of the second conjugation, where some semantic coherence was detected (possibly involving residual inchoative semantics),[9] verbs from other conjugations have not been found here to have more similar lexical semantics, which appears to support the exclusively morphological, that is, "(rhyzo)morphomic" nature of this lexical classification in Spanish. These findings are compatible with those of other researchers (e.g. Guzmán Naranjo [2020: 248] on Russian noun declensions), who also find that "on its own, semantics is not a very good predictor of inflection".

6 Conclusions

Over the last three decades, proponents and detractors of autonomous morphology have discussed whether languages can have (productive) grammatical structures or rules that span functionally incoherent domains, be these sets of paradigm cells (morphomes) or sets of lexemes (inflection classes). Most of these debates (often centered on Romance verb stem alternations) have relied on (i) the ability of researchers to qualitatively identify all features, values, and functions in paradigms, and (ii) a black-or-white conception of naturalness by which sets of cells (e.g. second person singular and third person singular) either are or are not a natural class, depending on

8 Although this would require additional investigation, it would not be unexpected if an exploration of lower-frequency verbs revealed something different, as a list of low-frequency verbs would contain many "productively derived" verb pairs like *escribir* 'write' and *re-escribir* 'rewrite'; *leer* 'read' and *re-leer* 'reread'; and *pensar* 'think' and *re-pensar* 'rethink'. Unlike high-frequency morphological derivatives (e.g. *conocer* 'know' and *re-conocer* 'recognize/admit'; *tener* 'have' and *re-tener* 'keep'; *pedir* 'ask' and *des-pedir* 'say goodbye'), these tend to be (or need to be) semantically compositional.

9 A frequent verbalizing suffix *-ecer* exists in Spanish that turns some adjectives and nouns into verbs (e.g. *pálido* 'pale' > *palidecer* 'become pale', *rojo* 'red' > *enrojecer 'become red'*, *noche* 'night' > *anochecer* 'become night', *flor* 'flower' > *florecer* 'blossom'). Although none of these straightforwardly derived verbs are amongst the 200 most frequent ones we analyze, some inchoative semantics, and a "group identity" of sorts might remain among verbs that continue the Latin inchoative: *conocer* '(come to) know', *aparecer* 'appear', *establecer* 'establish', *reconocer* 'recognize/admit', *nacer* 'be born' and *crecer* 'grow up'.

whether they do or do not share some abstract feature value(s) (e.g. +singular, –speaker) to the exclusion of other cells. Point (i) introduces a large subjective component into analyses, while (ii) imposes a binary taxonomization into a variable that is almost undoubtedly gradient (or at least very high-dimensional).

The best way forward must involve overcoming these limitations. Novel methods to measure the contextual similarity of words offer a promising avenue to probe on the architecture of paradigms and the lexicon. They can provide more quantitative and replicable evidence to any debate concerning syntactico-semantic natural classes. In this paper, we have assessed how (dis)similar different verbal word forms and lemmas are in Spanish, in order to quantify the (un)naturalness of those structures often described as morphomic (see e.g. Maiden 2018). Based on a corpus of over 1.4 billion words, we assessed the distributional-semantic similarity of verbs with same or different patterns of stem alternation, and from the same or different conjugations. We found that, whereas Spanish conjugations do not seem to have consistent semantic correlates (and could thus be quite fairly described as morphomic), stem alternations do. At the morphosyntactic level, the various word forms from the L, P, and F domains have all been found to involve higher similarity (and so naturalness) than randomly selected sets of cells of the same size. P has emerged as the most natural domain, and N as the least. At the lexical level, verbs with identical or similar patterns of stem alternation also tend to be significantly more similar semantically and/or syntactically.

We believe that these results increase our understanding of the (extra-morphological) functions of Romance verb stem alternations. Some have conceived these as diachronic "junk" (Lass 1990), that is, inherently useless remnants of former categories or sound changes, or maybe as useful only within the domain of morphology (by) itself (Aronoff 1994), for example, in relation to the Paradigm Cell Filling Problem (Ackerman and Malouf 2013; Blevins 2016). Here, by contrast, we found that Romance (specifically Spanish) verb stem alternations seem to be associated to more homogeneous lexical and inflectional domains than previously realized. In a way not unlike phonesthemes (Bergen 2004; Pimentel et al. 2019), they bind together lexemes and word forms that are comparatively similar semantically or syntactically, even when we cannot identify any one single meaning component that they all share exclusively. Even loose webs of semantic-cum-morphological similarities may be useful in the learning and processing of language in ways we are only beginning to understand.

Future research could involve experiments that investigate the cognitive advantages of (more) naturalness (see e.g. Saldana et al. 2022) or of abstract morphological properties and categories (see Veríssimo and Clahsen 2009). Assessing the weight of different extramorphological predictors in Romance stem alternations (Herce 2022a), the fit of different feature structure hypotheses to word-embedding similarities, or the distributional-semantic profile of stem alternations and conjugations in other languages would also be most welcome. We leave these for future research.

Supplementary Material

This article contains supplementary material (https://osf.io/eytk9/).

Acknowledgment: We would like to thank the editor of LV and an anonymous reviewer for their helpful comments.
Research funding: The second author expresses his gratitude for the support of grants from the French National Research Agency (ANR-20-CE27-0021).

References

Aalberse, Suzanne Pauline. 2007. The typology of syncretisms and the status of feature structure: Verbal paradigms across 355 Dutch dialects. *Morphology* 17(1). 109–149.
Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3). 429–464.

Allen, Nicholas J. 2008. Tetradic theory and the origin of human kinship systems. In Nicholas J. Allen, Hilary Callan, Robin Dunbar & Wendy James (eds.), *Early human kinship: From sex to social reproduction*, 96–112. Oxford: Blackwell.

Andersen, Henning. 2008. Naturalness and markedness. In Klaas Willems & Ludovic De Cuypere (eds.), *Naturalness and iconicity in language*, 101–119. Amsterdam: John Benjamins.

Anderson, Stephen R. 2013. Stem alternations in Swiss Rumantsch. In Silvio Cruschina, Martin Maiden & John Charles Smith (eds.), *The boundaries of pure morphology: Diachronic and synchronic perspectives*, 262–283. Oxford: Oxford Academic.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Cambridge, MA: MIT Press.

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity* 2019. 1–39.

Bentin, Shlomo & Laurie B. Feldman. 1990. The contribution of morphological and semantic relatedness to repetition priming at short and long lags: Evidence from Hebrew. *The Quarterly Journal of Experimental Psychology* 42(4). 693–711.

Bergen, Benjamin K. 2004. The psychological reality of phonaesthemes. *Language* 80(2). 290–311.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Bonami, Olivier & Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2). 173–196.

Booij, Geert. 1996. Inherent versus contextual inflection and the split morphology hypothesis. In Geert Booij & Jaap Marle (eds.), *Yearbook of morphology 1995*, 1–16. Amsterdam: Springer.

Boyé, Gilles & Gauvain Schalchli. 2019. Realistic data and paradigms: The paradigm cell finding problem. *Morphology* 29(2). 199–248.

Cardellino, Cristian. 2019. Spanish billion words corpus and embeddings. Available at: https://crscardellino.github.io/SBWCE/.

Chemla, Emmanuel, Toben H. Mintz, Savita Bernal & Anne Christophe. 2009. Categorizing words using "frequent frames": What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science* 12(3). 396–406.

Chuang, Yu-Ying, Dunstan Brown, R. Harald Baayen & Roger Evans. 2022. Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigm. *The Mental Lexicon* 17. 395–421.

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308.

Esher, Louise. 2012. *Future, conditional, and autonomous morphology in Occitan*. Oxford: University of Oxford Phd. Available at: https://ora.ox.ac.uk/objects/uuid:ba3acc5a-4474-4511-93c4-347bd2128b8d.

Esher, Louise. 2017. Morphome death and transfiguration in the history of French 1. *Journal of Linguistics* 53(1). 51–84.

Firth, John R. 1957. A synopsis of linguistic theory, 1930–1955. In John Firth (ed.), *Studies in linguistic analysis*, 1–32. Oxford: Basil Blackwell.

Guzmán Naranjo, Matías. 2020. Analogy, complexity and predictability in the Russian nominal inflection system. *Morphology* 30(3). 219–262.

Harbour, Daniel. 2016. *Impossible persons*. Cambridge, MA: MIT Press.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Herce, B. 2020a. On morphemes and morphomes: Exploring the distinction. *Word Structure* 13(1). 45–68.

Herce, Borja. 2020b. Stem alternations in Kiranti and their implications for the morphology–phonology interface. *Journal of Linguistics* 57(2). 321–363.

Herce, Borja. 2022a. Quantifying the importance of morphomic structure, semantic values, and frequency of use in Romance stem alternations. *Linguistics Vanguard* 8(1). 53–68.

Herce, Borja. 2022b. Stress and stem allomorphy in the Romance perfectum: Emergence, typology, and motivations of a symbiotic relation. *Linguistics* 60(4). 1103–1147.

Herce, Borja. 2023. *The typological diversity of morphomes: A cross-linguistic study of unnatural morphology*. Oxford: Oxford University Press.

Huyghe, Richard & Marine Wauquier. 2020. What's in an agent? *Morphology* 30(3). 185–218.

Kirschenbaum, Amit. 2021. Unsupervised induction of inflectional families. *Computer Speech & Language* 73. 101324.

Koontz-Garboden, Andrew. 2016. Thoughts on diagnosing morphomicity: A case study from Ulwa. In Ana R. Luís & Ricardo Bermúdez-Otero (eds.), *The morphome debate*, 228–247. Oxford: Oxford University Press.

Lass, Roger. 1990. How to do things with junk: Exaptation in language evolution. *Journal of Linguistics* 26(1). 79–102.

Luís, A. R. & R. Bermúdez-Otero. 2016. *The morphome debate*. Oxford: Oxford University Press.

Maiden, Martin. 2004. *When lexemes become allomorphs-on the genesis of suppletion*. Berlin/New York Berlin, New York: Walter de Gruyter.

Maiden, Martin. 2005. Morphological autonomy and diachrony. In Geert Booij & Jaap Marle (eds.), *Yearbook of morphology 2004*, 137–175. Dordrecht: Springer.

Maiden, Martin. 2017. Romansh allomorphy (again!). In Raffaella Zanuttini, Laurence Horn & Claire Bowern (eds.), *On looking into words (and beyond)*, 189–210. Berlin: Language Science Press.

Maiden, Martin. 2018. *The Romance verb: Morphomic structure and diachrony*. Oxford: Oxford University Press.

Malkiel, Yakov. 1966. Diphthongization, monophthongization, metaphony: Studies in their interaction in the paradigm of the Old Spanish -ir verbs. *Language* 42(2). 430–472.

McCarthy, Arya D., Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskij, Natalya Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra Jacobs, Ryan Cotterell, Mans Hulden & David Yarowsky. 2020. Unimorph 3.0: Universal morphology. *Proceedings of the 12th language resources and evaluation conference*, 3922–3931. European Language Resources Association. https://aclanthology.org/2020.lrec-1.483.

Miller, George A. & Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language & Cognitive Processes* 6(1). 1–28.
Mintz, Toben H. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition* 90(1). 91–117.
Nikolaev, Alexandre, Yu-Ying Chuang & R. Harald Baayen. 2022. A generating model for Finnish nominal inflection using distributional semantics. *The Mental Lexicon* 17(3). 368–394.
O'Neill, Paul. 2014. The morphome in constructive and abstractive models of morphology. *Morphology* 24(1). 25–70.
Pericliev, Vladimir & Raúl E. Valdés-Pérez. 1998. Automatic componential analysis of kinship semantics with a proposed structural solution to the problem of multiple models. *Anthropological Linguistics* 40(2). 272–317.
Pimentel, Tiago, Arya D. McCarthy, Damián E. Blasi, Brian Roark & Cotterell Ryan. 2019. Meaning to form: Measuring systematicity as information. https://arxiv.org/abs/1906.05906.
Radcliffe-Brown, Alfred Reginald. 1941. The study of kinship systems. *Journal of the Royal Anthropological Institute of Great Britain and Ireland* 71(1–2). 1–18.
Round, Erich R. 2015. Rhizomorphomes, meromorphomes and metamorphomes. In Matthew Baerman, Dunstan Brown & Greville G. Corbett (eds.), *Understanding and measuring morphological complexity*, 29–52. Oxford: Oxford University Press.
Saldana, Carmen, Borja Herce & Balthasar Bickel. 2022. More or less unnatural: Semantic similarity shapes the learnability and cross-linguistic distribution of unnatural syncretism in morphological paradigms. *Open Mind* 6. 1–28.
Smith, John Charles. 2013. The morphome as a gradient phenomenon: Evidence from Romance. In Silvio Cuschina, Martin Maiden & John Charles Smith (eds.), *The boundaries of pure morphology*, 247–261. Oxford: Oxford University Press.
Trommer, Jochen. 2016. A postsyntactic morphome cookbook. In Daniel Siddiqi & Heidi Harley (eds.), *Morphological metatheory*, 59–93. Amsterdam: John Benjamins.
Veríssimo, João & Harald Clahsen. 2009. Morphological priming by itself: A study of Portuguese conjugations. *Cognition* 112(1). 187–194.
Wyngaerd, Guido Vanden. 2018. The feature structure of pronouns: A probe into multidimensional paradigms. In Lena Baunaz, Liliane Haegeman, Karen De Clercq & Eric Lander (eds.), *Exploring nanosyntax*. Oxford: Oxford University Press.
Zwicky, Arnold. 1996. Syntax and phonology. In Keith Brown & Jim Miller (eds.), *Concise encyclopedia of syntactic theories*, 300–305. Oxford: Elsevier Science.